

GENERALIZED DOUBLE PARETO SHRINKAGE

Artin Armagan, David B. Dunson and Jaeyong Lee

SAS Institute Inc., Duke University and Seoul National University

Abstract: We propose a generalized double Pareto prior for Bayesian shrinkage estimation and inferences in linear models. The prior can be obtained via a scale mixture of Laplace or normal distributions, forming a bridge between the Laplace and Normal-Jeffreys' priors. While it has a spike at zero like the Laplace density, it also has a Student's t -like tail behavior. Bayesian computation is straightforward via a simple Gibbs sampling algorithm. We investigate the properties of the maximum a posteriori estimator, as sparse estimation plays an important role in many problems, reveal connections with some well-established regularization procedures, and show some asymptotic results. The performance of the prior is tested through simulations and an application.

Key words and phrases: Heavy tails, high-dimensional data, LASSO, maximum a posteriori estimation, relevance vector machine, robust prior, shrinkage estimation.

1. Introduction

There has been a great deal of work in shrinkage estimation and simultaneous variable selection in the frequentist framework. The LASSO of Tibshirani (1996) has drawn much attention to the area, particularly after the introduction of LARS (Efron et al. (2004)) due to its superb computational performance. There is a rich literature analyzing the LASSO and related approaches (Fu (1998), Knight and Fu (2000), Fan and Li (2001), Yuan and Lin (2005), Zhao and Yu (2006), Zou (2006), Zou and Li (2008)), with a number of articles considering asymptotic properties.

Bayesian approaches to the same problem became popular with the works of Tipping (2001) and Figueiredo (2003). By expressing Student's t priors for basis coefficients as scale mixtures of normals (West (1987)), and relying on type II maximum likelihood estimation (Berger (1985)), Tipping (2001) developed the relevance vector machine for sparse estimation in kernel regression. In this setting, however, exact sparsity comes with the price of forfeiting propriety of the posterior by driving the scale parameter of the Student's t distribution toward zero. In fact, driving both the scale parameter and the degrees of freedom to zero yields the so-called Normal-Jeffreys'

prior, $\pi(\theta) \propto 1/|\theta|$. The name emerges due to the fact that the hierarchy follows as $\theta \sim N(0, \tau)$, $\pi(\tau) \propto 1/\tau$, where the latter is the Jeffreys' prior on the prior variance of θ . Figueiredo (2003) proposed an expectation-maximization algorithm for maximum a posteriori estimation under Laplace and Normal-Jeffreys' priors, with estimates under the Laplace corresponding to the LASSO. The Normal-Jeffreys' prior leads to substantially improved performance with finite samples due to the property of strongly shrinking small coefficients to zero while minimally shrinking large coefficients due to the heavy tails; however, it has no meaning from an inferential aspect as it leads to an improper posterior.

A Bayesian LASSO was proposed by Park and Casella (2008) and Hans (2009). However, these procedures inherit the problem of over-shrinking large coefficients due to the relatively light tails of the Laplace prior. Strawderman-Berger priors (Strawderman (1971), Berger (1980)) have some desirable properties yet lack a simple analytic form. Recently proposed priors have been designed to have high density near zero and heavy tails without the impropriety problem of Normal-Jeffreys. The horseshoe prior of Carvalho, Polson, and Scott (2009, 2010) is induced through a carefully-specified mixture of normals, leading to such desirable properties as an infinite spike at zero and very heavy tails. They studied sparse shrinkage estimation properties of the horseshoe in a normal means problem. Griffin and Brown (2007, 2010) proposed an alternative class of hierarchical priors for shrinkage with some similarities to the prior we propose, but it lacks a simple analytic form that facilitates the study of some properties.

There is a need for alternative shrinkage priors that lead to sparse point estimates if desired, do not over-shrink coefficients that are not close to zero, facilitate straightforward computation even in large p cases, and result in a joint posterior distribution that does a good job of quantifying uncertainty. We propose the generalized double Pareto prior which independently finds mention in Cevher (2009). It has a simple analytic form, yields a proper posterior, and possesses such appealing properties as a spike at zero, Student's t -like tails, and a simple characterization as a scale mixture of normals that leads to a straightforward Gibbs sampler for posterior inferences. We consider both fully Bayesian and frequentist penalized likelihood approaches based on this prior. We show that the induced penalty in the regularization framework yields a consistent thresholding rule having the continuity property in the orthogonal case, with a simple expectation-maximization algorithm described for sparse estimation in non-orthogonal

cases. In another independent work motivated by applications to genome wide associations studies, Lee et al. (2011) consider a generalized t prior (McDonald and Newey (1988)) that includes the generalized double Pareto as a special case. Similarities to previous work are limited and our contributions beyond them are (i) the formal introduction of a generalized Pareto density, thresholded and folded at zero, as a shrinkage prior in Bayesian analysis, (ii) the scale mixture representation of the generalized double Pareto in Proposition 1 which is central to our work, (iii) its connection to the Laplace and Normal-Jeffreys' priors as limiting cases in Proposition 2, (iv) the resulting fully conditional posteriors in a linear regression setting along with a simple Gibbs sampling procedure, (v) a detailed discussion on the hyper-parameters α and η and their treatment, along with the incorporation of a griddy sampling scheme into the Gibbs sampler, (vi) a detailed analysis of the induced penalty by the generalized double Pareto prior and the properties of the resulting thresholding rule, (vii) an explicit analytic form for the maximum a posteriori estimator in orthogonal cases, (viii) an expectation-maximization procedure to obtain the maximum a posteriori estimate in non-orthogonal cases using the normal mixture representation, (ix) the one-step estimator (Zou and Li (2008)) resulting from the Laplace mixture representation, revealing the connection of the resulting procedure to the adaptive LASSO of Zou (2006), and (x) the oracle properties of the resulting estimators.

2. Generalized Double Pareto Prior

The generalized double Pareto density is

$$f(\theta|\xi, \alpha) = \frac{1}{2\xi} \left(1 + \frac{|\theta|}{\alpha\xi}\right)^{-(\alpha+1)}, \quad (2.1)$$

where $\xi > 0$ is a scale parameter and $\alpha > 0$ is a shape parameter. In contrast to (2.1), the generalized Pareto density of Pickands (1975) is parametrized in terms of a location parameter $\mu \in \mathbb{R}$, a scale parameter $\xi > 0$, and a shape parameter $\alpha \in \mathbb{R}$ as

$$f(\theta|\xi, \alpha, \mu) = \frac{1}{\xi} \left(1 + \frac{\theta - \mu}{\alpha\xi}\right)^{-(\alpha+1)}, \quad (2.2)$$

with $\theta \geq \mu$ for $\alpha > 0$ and $\mu \leq \theta \leq \mu - \xi\alpha$ for $\alpha < 0$. The mean and variance for the generalized Pareto distribution are $\mathbb{E}(\theta) = \mu + \xi/(1 - 1/\alpha)$ for $\alpha \notin [0, 1]$ and $\mathbb{V}(\theta) = \xi^2(1 - 1/\alpha)^{-2}(1 - 2/\alpha)^{-1}$ for $\alpha \notin [0, 2]$. If we let $\mu = 0$, (2.2) becomes an

exponential density as $\alpha \rightarrow \infty$ with mean ξ and variance ξ^2 .

To modify the generalized Pareto density to be a shrinkage prior, we let $\mu = 0$ and reflect the positive part about the origin, assuming $\alpha > 0$, for a density that is symmetric about zero. The mean and variance for the generalized double Pareto distribution are $\mathbb{E}(\theta) = 0$ for $\alpha > 1$ and $\mathbb{V}(\theta) = 2\xi^2\alpha^2(\alpha - 1)^{-1}(\alpha - 2)^{-1}$ for $\alpha > 2$. The dispersion is controlled by ξ and α , with α controlling the tail heaviness and $\alpha = 1$ corresponding to Cauchy-like tails and no finite moments.

Figure 2.1 compares the density in (2.1) to Cauchy and Laplace densities for the special case $\xi = \alpha = 1$, so that $f(\theta) = 1/\{2(1 + |\theta|)^2\}$. We refer to this form as the standard double Pareto. Near zero, the standard double Pareto resembles the Laplace density, suggesting similar sparse shrinkage properties of small coefficients in maximum a posteriori estimation. It also has Cauchy-like tails, which is appealing in avoiding over-shrinkage away from the origin. This is illustrated in Figure 2.1(a). Figure 2.1(b) illustrates how the density in (2.1) changes for different values of ξ and α .

Prior (2.1) can be represented as a scale mixture of normal distributions leading to computational simplifications. As shorthand notation, let $\theta \sim \text{GDP}(\xi, \alpha)$ denote that θ has density (2.1).

Proposition 1. *Let $\theta \sim N(0, \tau)$, $\tau \sim \text{Exp}(\lambda^2/2)$, and $\lambda \sim \text{Ga}(\alpha, \eta)$, where $\alpha > 0$ and $\eta > 0$. The resulting marginal density for θ is $\text{GDP}(\xi = \eta/\alpha, \alpha)$.*

Proposition 1 reveals a relationship between the prior in (2.1) and the prior of Griffin and Brown (2007), with the difference being that Griffin and Brown (2007) place a mixing distribution on λ^2 leading to a marginal density on θ with no simple analytic form.

In Proposition 2 we show that the prior in (2.1) forms a bridge between two limiting cases – Laplace and Normal-Jeffreys’ priors.

Proposition 2. *Given the representation in Proposition 1, $\theta \sim \text{GDP}(\xi = \eta/\alpha, \alpha)$ implies*

1. $f(\theta) \propto 1/|\theta|$ for $\alpha = 0$ and $\eta = 0$,
2. $f(\theta|\lambda') = (\lambda'/2) \exp(-\lambda'|\theta|)$ for $\alpha \rightarrow \infty$, $\alpha/\eta = \lambda'$ and $0 < \lambda' < \infty$.

Proof. For the first item, setting $\alpha = \eta = 0$ implies placing a Jeffreys’ prior on λ , $\pi(\lambda) \propto 1/\lambda$. Integration over λ yields $\pi(\tau) \propto 1/\tau$, which implies the Normal-Jeffreys’ prior on θ . For the second item, notice that $\pi(\lambda) = \delta(\lambda - \lambda')$, where $\delta(\cdot)$ denotes the

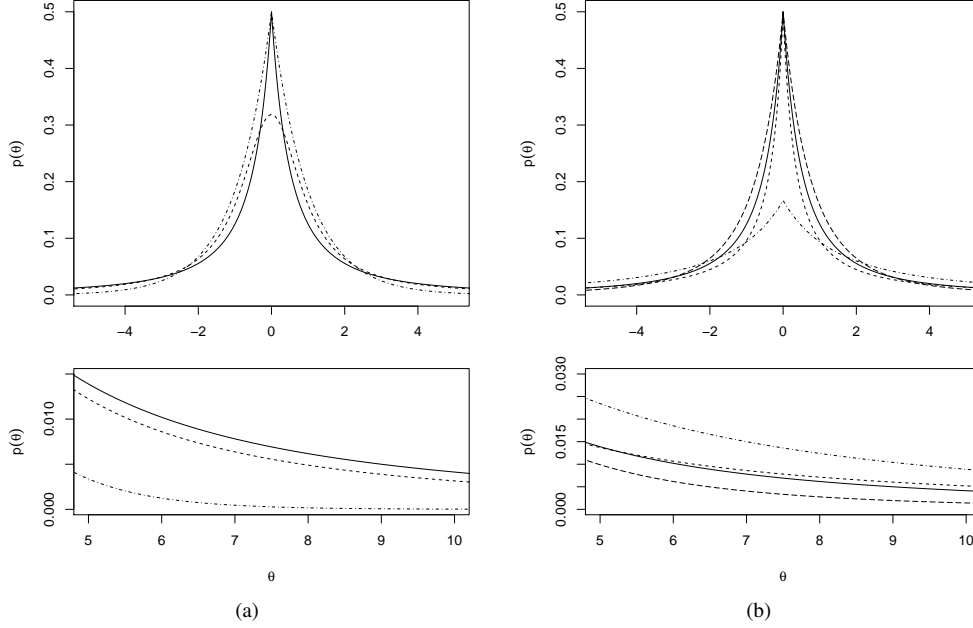


Figure 2.1: (a) Probability density functions for standard double Pareto (solid line), standard Cauchy (dashed line) and Laplace (dot-dash line) ($\lambda = 1$) distributions. (b) Probability density functions for the generalized double Pareto with (ξ, α) values of (1, 1) (solid line), (0.5, 1) (dashed line), (1, 3) (long-dashed line), and (3, 1) (dot-dash line).

Dirac delta function, since $\lim_{\alpha \rightarrow \infty} \lim_{\alpha/\eta \rightarrow \lambda'} \mathbb{E}(\lambda) = \lambda'$ and $\lim_{\alpha \rightarrow \infty} \lim_{\alpha/\eta \rightarrow \lambda'} \mathbb{V}(\lambda) = 0$. Thus, $\int_0^\infty (\lambda/2) \exp(-\lambda|\theta|) \delta(d\lambda) = (\lambda'/2) \exp(-\lambda'|\theta|)$. \square

As noted in Polson and Scott (2010), if $\pi(\tau)$ has exponential or lighter tails, observations are shrunk towards zero by some non-diminishing amount, regardless of size. This phenomenon is well-understood and commonly observed in estimation under the Laplace prior, where an exponential density mixes a normal density. The higher-level mixing (over λ) in Proposition 1 allows $\pi(\tau)$ to have heavier tails, remedying the unwanted bias.

As α grows, the density becomes lighter tailed, more peaked and the variance becomes smaller, while as η grows, the density becomes flatter and the variance increases. Hence if we increase α , we may cause unwanted bias for large signals, though causing stronger shrinkage for noise-like signals; if we increase η we may lose the ability to

shrink noise-like signals, as the density is not as pronounced around zero; and finally, if we increase α and η at the same rate, the variance remains constant but the tails become lighter, converging to a Laplace density in the limit. This leads to over-shrinking of coefficients that are away from zero. As a typical default specification for the hyperparameters, one can take $\alpha = \eta = 1$. This choice leads to Cauchy-like tail behavior, which is well-known to have desirable Bayesian robustness properties.

To motivate this default choice, we assess the behavior of the prior shrinkage factor $\kappa = 1/(1 + \tau) \in (0, 1)$, where $\theta \sim N(0, \tau)$ is the parameter of interest (Carvalho et al. (2010)). As $\kappa \rightarrow 0$, the prior imposes no shrinkage, while as $\kappa \rightarrow 1$ it has a strong pull towards zero. The generalized double Pareto distribution implies a prior $\pi(\kappa)$ on κ upon integration over λ in Proposition 1. For the standard double Pareto, this is

$$\pi(\kappa) = \frac{1}{2(1 - \kappa)^2} \left[\frac{\sqrt{\pi} \exp \left\{ \frac{\kappa}{2(1 - \kappa)} \right\} \operatorname{Erfc} \left\{ \sqrt{\frac{\kappa}{2(1 - \kappa)}} \right\}}{\sqrt{2\kappa(1 - \kappa)}} - 1 \right],$$

where $\operatorname{Erfc}(\cdot)$ denotes the complementary error function. In Figure 2.2, we compare $\pi(\kappa)$ under the standard double Pareto, Strawderman-Berger, horseshoe, and Cauchy priors, which may all be considered default choices. The priors behave similarly for $\kappa \approx 0$, implying similar tail behavior. The behavior of $\pi(\kappa)$ for $\kappa \approx 1$ governs the strength of shrinkage of small signals. As $\kappa \rightarrow 1$, $\pi(\kappa)$ tends towards zero for the Cauchy, implying weak shrinkage, while $\pi(\kappa)$ is unbounded for the horseshoe, suggesting a strong pull towards zero for small signals. The Strawderman-Berger and standard double Pareto priors are a compromise between these extremes, with $\pi(\kappa)$ bounded for $\kappa \rightarrow 1$ in both cases. The standard double Pareto assigns higher density close to one than the Strawderman-Berger prior, and has the advantage of a simple analytic form over the Strawderman-Berger and horseshoe priors.

Of course it is best to adjust α and η according to any available prior information pertaining to the sparsity structure of the estimated vector. For general $\alpha > 0$ and $\eta > 0$

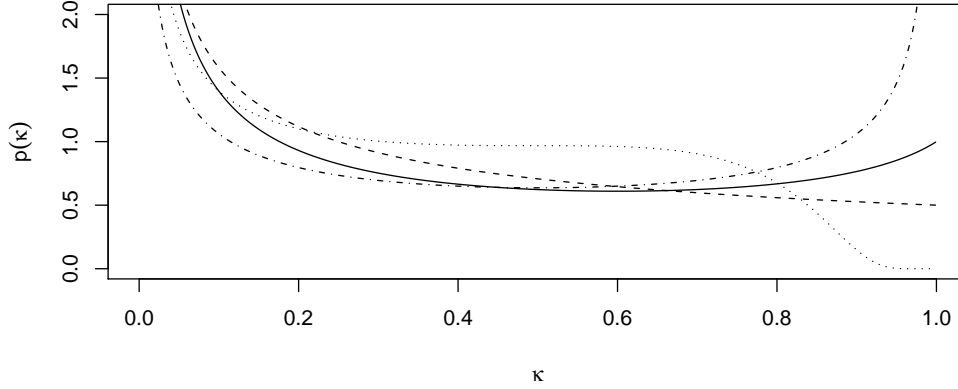


Figure 2.2: Prior density of κ implied by the standard double Pareto prior (solid line), Strawderman–Berger prior (dashed line), horseshoe prior (dot-dash line) and standard Cauchy prior (dotted line).

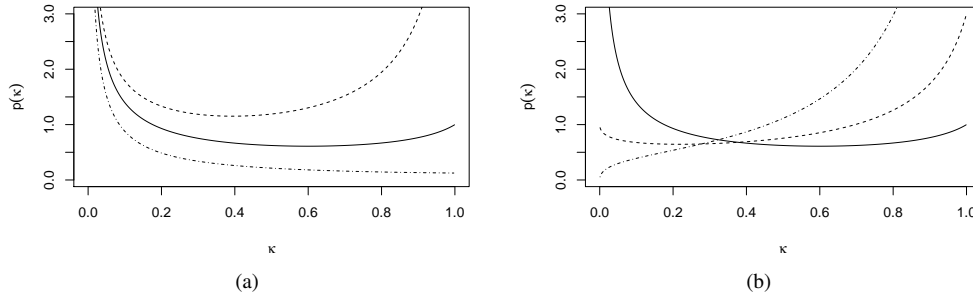


Figure 2.3: Prior density of κ (a) when $\alpha = 1$ and $\eta = 0.5$ (dashed), $\eta = 1$ (solid), $\eta = 2$ (dot-dash) (b) when $\eta = 1$ and $\alpha = 1$ (solid), $\alpha = 2$ (dashed), $\alpha = 3$ (dot-dash).

values, the prior on κ is

$$\begin{aligned}
 \pi(\kappa|\alpha, \eta) &= \frac{2^{\alpha/2-1} \eta^\alpha \kappa^{(\alpha-1)/2} (1-\kappa)^{-(\alpha+3)/2}}{\Gamma(\alpha)} \\
 &\times \left\{ \left(\frac{1}{\kappa} - 1 \right)^{1/2} \Gamma\left(\frac{\alpha}{2} + 1\right) {}_1F_1\left(\frac{\alpha}{2} + 1, \frac{1}{2}, \frac{\eta^2 \kappa}{2(1-\kappa)}\right) \right. \\
 &\quad \left. - \sqrt{2} \eta \Gamma\left(\frac{\alpha+3}{2}\right) {}_1F_1\left(\frac{\alpha+3}{2}, \frac{3}{2}, \frac{\eta^2 \kappa}{2(1-\kappa)}\right) \right\}, \quad (2.3)
 \end{aligned}$$

where ${}_1F_1$ denotes the confluent hypergeometric function. Note that $\pi(\kappa|\alpha, \eta)$ takes a “horseshoe” shape when $\alpha = \eta = 0$. Carvalho, Polson, and Scott (2010) show that $\pi(\kappa) \propto \kappa^{-1}(1-\kappa)^{-1}$ implies a Normal-Jeffreys’ prior on θ , which can also be observed by setting $\alpha = \eta = 0$ in (2.3) in conjunction with Proposition 1. Hence $\pi(\kappa|\alpha, \eta)$ is unbounded at $\kappa = 1$ forcing $\pi(\theta|\alpha, \eta)$ to be unbounded at 0 only if $\eta = 0$. The effects of α and η are now observed with better clarity from Figure 2.3. As η increases, less and less density is assigned to the neighborhood of $\kappa \approx 1$, repressing shrinkage. On the other hand, increasing α values place more and more density in the neighborhood of $\kappa \approx 1$ promoting further shrinkage. This notion is later reinforced by Proposition 3, such that the prior induces a thresholding rule under maximum a posteriori estimation if $\eta < 2\sqrt{\alpha + 1}$. Hence, we need to carefully pick these hyper-parameters, in particular α , as there is a trade-off between the magnitude of shrinkage and tail robustness.

3. Bayesian Inference in Linear Models

Consider the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where \mathbf{y} is an n -dimensional vector of responses, \mathbf{X} is the $n \times p$ design matrix and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$. Letting $\beta_j|\sigma \sim \text{GDP}(\xi = \sigma\eta/\alpha, \alpha)$ independently for $j = 1, \dots, p$,

$$\pi(\boldsymbol{\beta}|\sigma) = \prod_{j=1}^p \frac{1}{2\sigma\eta/\alpha} \left(1 + \frac{1}{\alpha} \frac{|\beta_j|}{\sigma\eta/\alpha} \right)^{-(\alpha+1)}. \quad (3.1)$$

From Proposition 1, this prior is equivalent to $\beta_j|\sigma \sim N(0, \sigma^2\tau_j)$, with $\tau_j \sim \text{Exp}(\lambda_j^2/2)$ and $\lambda_j \sim \text{Ga}(\alpha, \eta)$. We place the Jeffreys’ prior on the error variance, $\pi(\sigma) \propto 1/\sigma$.

Using the scale mixture of normals representation, we obtain a simple data augmentation Gibbs sampler having the conditional posteriors $(\boldsymbol{\beta}|\sigma^2, \mathbf{T}, \mathbf{y}) \sim N\{(\mathbf{X}'\mathbf{X} + \mathbf{T}^{-1})^{-1}\mathbf{X}'\mathbf{y}, \sigma^2(\mathbf{X}'\mathbf{X} + \mathbf{T}^{-1})^{-1}\}$, $(\sigma^2|\boldsymbol{\beta}, \mathbf{T}, \mathbf{y}) \sim \text{IG}\{(n+p)/2, (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/2 + \boldsymbol{\beta}'\mathbf{T}^{-1}\boldsymbol{\beta}/2\}$, $(\lambda_j|\beta_j, \sigma^2) \sim \text{Ga}(\alpha + 1, |\beta_j|/\sigma + \eta)$, $(\tau_j^{-1}|\beta_j, \lambda_j, \sigma^2) \sim \text{Inv-Gauss}\{\mu = (\lambda_j^2\sigma^2/\beta_j^2)^{1/2}, \rho = \lambda_j^2\}$, where $\mathbf{T} = \text{diag}(\tau_1, \dots, \tau_p)$ and Inv-Gauss denotes the inverse Gaussian distribution with location and scale parameters μ and ρ . In our experience, this Gibbs sampler is efficient with fast rates of convergence and mixing.

In the absence of any prior information on α and η , one may either set them to their default values or, as an alternative, choose hyper-priors to allow the data to inform about the values of α and η . We use $\pi(\alpha) = 1/(1+\alpha)^2$ and $\pi(\eta) = 1/(1+\eta)^2$ to correspond to generalized Pareto hyper-priors with location parameter 0, scale parameter 1 and shape

parameter 1. The median value of the resulting distribution for α and η is 1, centered at the default choices suggested earlier, while the mean and variance do not exist.

For sampling purposes, let $a = 1/(1+\alpha)$ and $e = 1/(1+\eta)$. These transformations suggest a uniform prior on a and e in $(0, 1)$ given the generalized Pareto priors on α and η . Consequently, the conditional posteriors for a and e are

$$\begin{aligned}\pi(a|\boldsymbol{\beta}, \eta) &\propto \left(\frac{1-a}{a}\right)^p \prod_{j=1}^p \left(1 + \frac{|\beta_j|}{\sigma\eta}\right)^{-1/a}, \\ \pi(e|\boldsymbol{\beta}, \alpha) &\propto \left(\frac{e}{1-e}\right)^p \prod_{j=1}^p \left\{1 + e \frac{|\beta_j|}{\sigma(1-e)}\right\}^{-(\alpha+1)}.\end{aligned}$$

We propose the embedded griddy Gibbs (Ritter and Tanner (1992)) sampling scheme:

- i. Form a grid of m points $a^{(1)}, \dots, a^{(m)}$ in the interval $(0, 1)$.
- ii. Calculate $w^{(k)} = \pi(a^{(k)}|\boldsymbol{\beta}, \eta)$.
- iii. Normalize the weights, $w_N^{(k)} = w^{(k)} / \sum_{k=1}^m w^{(k)}$.
- iv. Draw a sample from the set $\{a^{(1)}, \dots, a^{(m)}\}$ with probabilities $\{w_N^{(1)}, \dots, w_N^{(m)}\}$, and set $\alpha = 1/a - 1$ to be used at the current iteration of the Gibbs sampler.

Repeat the same procedure for e and obtain a random draw for η . We also experiment with fixing η as 1 while treating α as unknown. In this case, the prior variance of $\boldsymbol{\beta}|\sigma^2$ is determined by α .

In what follows we establish the ties between the Bayesian approach we have taken and some frequentist regularization approaches. The simple analytic structure of the generalized double Pareto prior facilitates analyses while its hierarchical formulation leads to straight-forward computation.

4. Sparse Maximum a Posteriori Estimation

The generalized double Pareto distribution can be used not only as a prior in a Bayesian analysis, but also to induce a sparsity-favoring penalty in regularized least squares:

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \left\{ \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p p(|\beta_j|) \right\}, \quad (4.1)$$

where \mathbf{X} is initially assumed to have orthonormal columns and $p(\cdot)$ denotes the penalty function implied by the prior on the regression coefficients. Following Fan and Li (2001), let $\hat{\beta} = \mathbf{X}'\mathbf{y}$, and denote the minimization problem in (4.1) for a component of β as

$$\tilde{\beta}_j = \arg \min_{\beta_j} \left\{ \frac{1}{2} (\hat{\beta}_j - \beta_j)^2 + \sigma^2 p(|\beta_j|) \right\}, \quad (4.2)$$

with the penalty function $p(|\beta_j|) = (\alpha + 1) \log(\sigma\eta + |\beta_j|)$ that simply retains the term in $-\log \pi(\beta_j|\alpha, \eta)$ that depends on β_j .

From Fan and Li (2001), a good penalty function should result in an estimator that is (i) nearly unbiased when the true unknown parameter is large, (ii) a thresholding rule that automatically sets small estimated coefficients to zero to reduce model complexity, and (iii) continuous in data ($\hat{\beta}_j$) to avoid instability in model prediction. In the following, we show that the penalty function induced by prior (3.1) may achieve these properties.

4.1. Near-unbiasedness

The first order derivative of (4.2) with respect to β_j is $\text{sgn}(\beta_j)\{|\beta_j| + \sigma^2 p'(|\beta_j|)\} - \hat{\beta}_j = \text{sgn}(\beta_j)\{|\beta_j| + \sigma^2(\alpha + 1)/(\sigma\eta + |\beta_j|)\} - \hat{\beta}_j$, where $p'(|\beta_j|) = \partial p(|\beta_j|)/\partial |\beta_j|$ is the term causing bias in estimation. Although it is appealing to introduce bias in small coefficients to reduce the mean squared error and model complexity, it is also desirable to limit the shrinkage of large coefficients with $p'(|\beta_j|) \rightarrow 0$ as $|\beta_j| \rightarrow \infty$. In addition, it is desirable for $p'(|\beta_j|)$ to approach zero rapidly, implying shrinkage, and the associated introduction of bias rapidly decreases as coefficients get further away from zero. In fact, the rate of convergence of $p'(|\beta_j|)$ to zero is of the same order under the generalized double Pareto and Normal-Jeffreys' priors, with $\lim_{|\beta_j| \rightarrow \infty} \{(\alpha + 1)/(\sigma\eta + |\beta_j|)\} / \{1/|\beta_j|\} = \alpha + 1$. As α controls the tail heaviness in the generalized double Pareto prior, with lighter tails for larger values of α , convergence of the ratio to $(\alpha + 1)$ is intuitive. In the case of LASSO, the bias, $p'(|\beta_j|)$, remains constant regardless of $|\beta_j|$, which can also be observed in Figure 4.4(b).

4.2. Sparsity

As noted in Fan and Li (2001), a sufficient condition for the resulting estimator to be a thresholding rule is that the minimum of the function $|\beta_j| + \sigma^2 p'(|\beta_j|)$ is positive.

Proposition 3. *Under the formulation in Proposition 1, prior (3.1) implies a penalty yielding an estimator that is a thresholding rule if $\eta < 2\sqrt{\alpha + 1}$.*

This result is obtained by finding the minimum of $|\beta_j| + \sigma^2 p'(|\beta_j|)$ and taking it greater than zero. The thresholding is a direct consequence of the fact that when $|\hat{\beta}_j| < \min_{\beta_j} \{|\beta_j| + \sigma^2(\alpha+1)/(\sigma\eta + |\beta_j|)\}$, which requires that $\min_{\beta_j} \{|\beta_j| + \sigma^2 p'(|\beta_j|)\} > 0$, the derivative of (4.2) is positive for all positive β_j and negative for all negative β_j . In this case, the penalized least squares estimator is zero. When $|\hat{\beta}_j| > \min_{\beta_j} \{|\beta_j| + \sigma^2(\alpha+1)/(\sigma\eta + |\beta_j|)\}$, two roots may exist. The larger one (in absolute value) or zero is the penalized least squares estimator. To elaborate more on this, the root(s) may exist for $\text{sgn}(\beta_j)\{|\beta_j| + \sigma^2 p'(|\beta_j|)\} - \hat{\beta}_j = 0$ only when $|\hat{\beta}_j| > \min_{\beta_j} \{|\beta_j| + \sigma^2 p'(|\beta_j|)\}$. A helpful illustration is Figure 3 of Fan and Li (2001).

4.3. Continuity

Continuity in data is important if an estimator is to avoid instabilities in prediction. As in Breiman (1996), “a regularization procedure is unstable if a small change in data can make large changes in the regularized estimator”. Discontinuities in the thresholding rule may result in inclusion or dismissal of a signal with minor changes in the data used (see Figure 4.4(b)). Hard-thresholding, the “usual” variable selection, is an unstable procedure, while ridge and LASSO estimates are considered stable.

A necessary and sufficient condition for continuity is that the minimum of the function $|\beta_j| + \sigma^2 p'(|\beta_j|)$ is at zero (Fan and Li (2001)). For our prior, the minimum of this function is obtained at $|\beta_j| = \sigma(\sqrt{\alpha+1} - \eta)$. Therefore $\eta = \sqrt{\alpha+1}$ yields an estimator with this property.

Proposition 4. *Under the formulation in Proposition 1, a subfamily of prior (2.1) with $\eta = \sqrt{\alpha+1}$ yields an estimator with the continuity property.*

In this particular case, the penalized likelihood estimator is set to zero if $|\hat{\beta}_j| \leq \sigma\sqrt{\alpha+1}$. When $|\hat{\beta}_j| > \sigma\sqrt{\alpha+1}$,

$$\tilde{\beta}_j = \begin{cases} \frac{\hat{\beta}_j - \sigma\sqrt{\alpha+1} + \{\hat{\beta}_j^2 + 2\hat{\beta}_j\sigma\sqrt{\alpha+1} - 3\sigma^2(\alpha+1)\}^{1/2}}{2} & \hat{\beta}_j > 0, \\ \frac{\hat{\beta}_j + \sigma\sqrt{\alpha+1} - \{\hat{\beta}_j^2 - 2\hat{\beta}_j\sigma\sqrt{\alpha+1} - 3\sigma^2(\alpha+1)\}^{1/2}}{2} & \hat{\beta}_j < 0. \end{cases} \quad (4.3)$$

As can be observed in Figure 4.4(a), ensuring continuity by letting $\eta = \sqrt{\alpha+1}$ creates a trade-off between sparsity and tail-robustness. As the thresholding region becomes wider, the larger values are penalized further, yet not nearly at the level of LASSO.

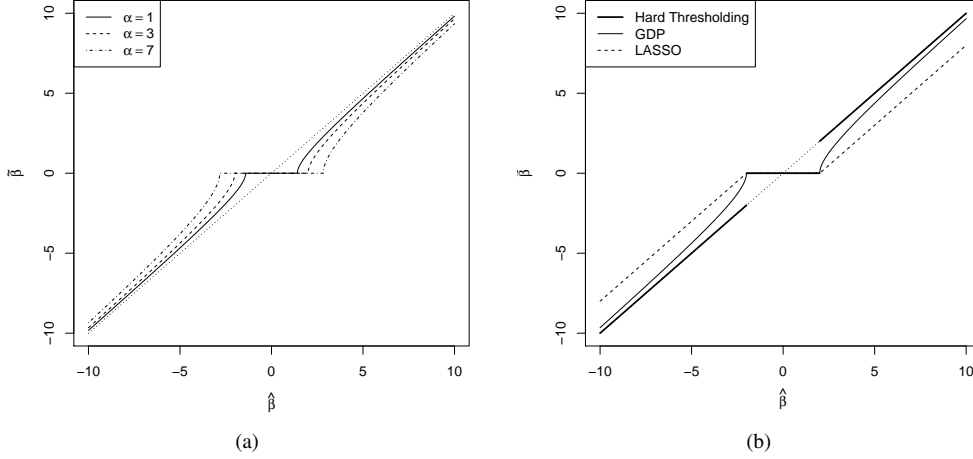


Figure 4.4: Thresholding functions for (a) generalized double Pareto prior with $\eta = \sqrt{\alpha + 1}$, $\alpha = \{1, 3, 7\}$, (b) Hard thresholding, generalized double Pareto prior with $\eta = 2$, $\alpha = 3$ and LASSO with $\sigma = 1$.

4.4. Maximum a Posteriori Estimation via Expectation-Maximization

We assume a normal likelihood to formulate the procedure for non-orthogonal linear regression. Estimation is carried out via the expectation-maximization (EM) algorithm.

4.4.1. Exploiting the Normal Mixture Representation

We take the expectation of the log-posterior with respect to the conditional posterior distributions of $(\tau_j^{-1}|\beta_j^{(k)}, \lambda_j, \sigma^{2(k)})$ and $(\lambda_j|\beta_j^{(k)}, \sigma^{2(k)})$ at the k th step, then maximize with respect to β_j and σ^2 to get the values for the $(k + 1)$ th step.

• *E-step:*

$$\begin{aligned}
 & - \left(\frac{n+p}{2} + 1 \right) \log \sigma^2 - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \\
 & - \frac{1}{2\sigma^2} \sum_{j=1}^p \beta_j^2 \underbrace{\left\{ \frac{(\alpha+1)\sigma^{2(k)}}{|\beta_j^{(k)}|(|\beta_j^{(k)}| + \sigma^{(k)}\eta)} \right\}}_{d_j^{(k)}}
 \end{aligned}$$

- *M-step*: Letting $\mathbf{D}^{(k)} = \text{diag}(d_1^{(k)}, \dots, d_p^{(k)})$, we have

$$\begin{aligned}\boldsymbol{\beta}^{(k+1)} &= (\mathbf{X}'\mathbf{X} + \mathbf{D}^{(k)})^{-1}\mathbf{X}'\mathbf{y}, \\ \sigma^{2(k+1)} &= \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k+1)})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k+1)}) + \boldsymbol{\beta}^{(k+1)'}\mathbf{D}^{(k)}\boldsymbol{\beta}^{(k+1)}}{n + p + 2}.\end{aligned}$$

We refer to this estimator as GDP(MAP).

4.4.2. Exploiting the Laplace Mixture Representation and the One-step Estimator

In the proof of Proposition 1, the integration over $\boldsymbol{\tau}$ leads to a Laplace mixture representation of the prior. Since the mixing distribution of the Laplace is a known distribution the required expectation is obtained with ease, resulting in the maximization step,

$$\begin{aligned}\boldsymbol{\beta}^{(k+1)} &= \\ \arg \max_{\boldsymbol{\beta}} &\left\{ -\frac{1}{2\sigma^{2(k)}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \frac{1}{\sigma^{(k)}} \sum_{j=1}^p |\beta_j| \left(\frac{\alpha + 1}{|\beta_j^{(k)}|/\sigma^{(k)} + \eta} \right) \right\}, \\ \sigma^{2(k+1)} &= \frac{b^2 - 2ac - \sqrt{b^4 - 4acb^2}}{2a^2},\end{aligned}\tag{4.4}$$

where $a = -(n + p + 2)$, $b = (\alpha + 1) \sum_j |\beta_j^{(k+1)}| / (|\beta_j^{(k)}|/\sigma^{(k)} + \eta)$, and $c = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k+1)})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(k+1)})$. The component-specific multiplier on $|\beta_j|$ is obtained from the expectation of λ_j with respect to its conditional posterior distribution, $\pi(\lambda_j|\beta_j, \sigma^2)$. Similar results to (4.4) are in Candès, Wakin and Boyd (2008), Cevher (2009), and Garrigues (2009).

An intuitive relationship to the adaptive LASSO of Zou (2006) and the one-step sparse estimator of Zou and Li (2008) can be seen via the Laplace mixture representation. As a computationally fast alternative to estimating the exact mode via the above EM algorithm, we can obtain a “one-step estimator” and exploit the LARS algorithm as in Zou and Li (2008). The one-step estimator is

$$\boldsymbol{\beta}^{(1)} = \arg \min_{\boldsymbol{\beta}} \left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \alpha^\dagger \sum_{j=1}^p \frac{|\beta_j|}{|\beta_j^{(0)}| + \eta^\dagger} \right\}, \tag{4.5}$$

with $\alpha^\dagger = 2\sigma^{2(0)}(\alpha + 1)$ and $\eta^\dagger = \sigma^{(0)}\eta$. This estimator resembles the adaptive LASSO. The LARS algorithm can be used to obtain $\beta^{(1)}$ very quickly. We refer to this estimator as GDP(OS).

Remark 1. For $\eta^\dagger = 0$, the GDP(OS) solution path for varying α^\dagger is identical to the adaptive LASSO solution path with $\gamma = 1$ (see (4) in Zou (2006)) using identical $\beta^{(0)}$.

Remark 2. GDP(OS) forms a bridge between the LASSO and the adaptive LASSO: as $\eta^\dagger \rightarrow \infty$ and $\alpha^\dagger/\eta^\dagger \rightarrow \lambda^\dagger < \infty$, GDP(OS) gives the LASSO solution with penalty parameter λ^\dagger .

We derive the GDP(OS) estimator only to reveal a close connection with the adaptive LASSO of Zou (2006) and do not use it in our experiments.

4.4.3. Normal vs. Laplace Representations in Computation

As pointed out by an anonymous referee, it is appropriate to compare the convergence behavior of the EM algorithms that exploit different mixture representations. We generated $n = \{200, 400, 600, 800, 1000\}$ observations from $y_i = \mathbf{x}_i' \beta^* + \epsilon_i$, where the x_{ij} were independent standard normals for $p = \{20, 40, 60, 80, 100\}$, $\epsilon_i \sim N(0, \sigma^2)$, and $\sigma = 3$. We set the first $p/4$ components of β^* to be 1 and the rest to 0. For each (n, p) combination we simulated 100 data sets and ran the EM algorithms obtained from normal and Laplace scale mixture representations. Figure 4.5 illustrates the number of iterations taken by the two algorithms until $\|\beta^{(k+1)} - \beta^{(k)}\|_2 < 10^{-6}$. As expected, the convergence under the Laplace mixture representation was much faster with the intermediary mixing parameter τ_j integrated out rather than using the expectation step in the EM algorithm.

4.5. Oracle Properties

Following Zou (2006) and Zou and Li (2008), we show that the GDP(MAP) and GDP(OS) estimators possess oracle properties. Relaxing the normality assumption on the error term leads to two conditions for Theorem 2 and Theorem 3.

(A1) $y_i = \mathbf{x}_i' \beta^* + \epsilon_i$ where $\epsilon_1, \dots, \epsilon_n$ are independent and identically distributed with mean 0 and variance σ^2 .

(A2) $\frac{1}{n} \mathbf{X}' \mathbf{X} \rightarrow \mathbf{C}$, where \mathbf{C} is a positive definite matrix.

In what follows, $\mathcal{A} = \{j : \beta_j^* \neq 0, j = 1, \dots, p\}$, $\beta_{\mathcal{A}}$ retains the entries of β indexed by \mathcal{A} , and $\mathbf{C}_{\mathcal{A}}$ retains the rows and columns of \mathbf{C} indexed by \mathcal{A} .

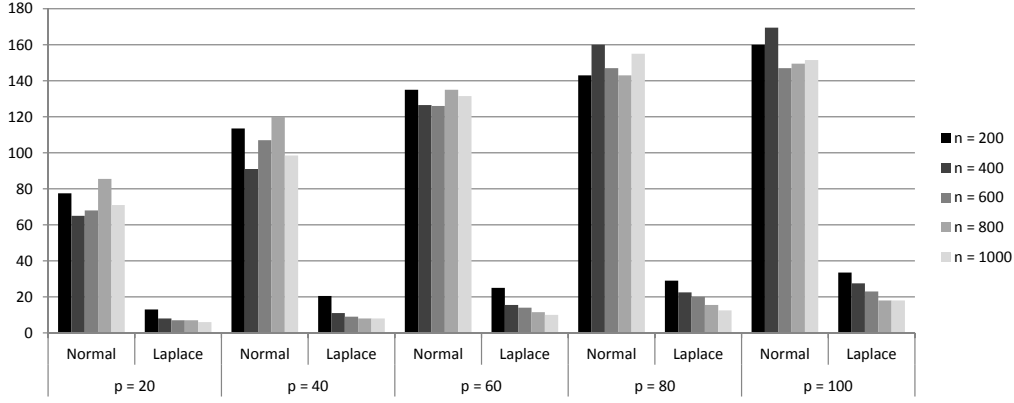


Figure 4.5: Number of iterations until convergence of the EM algorithms under normal and Laplace representations.

Theorem 1. *Let*

$$\beta_n^{(\infty)} = \arg \min_{\beta} \left\{ (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \alpha'_n \sum_{j=1}^p \log(|\beta_j| + \eta'_n) \right\}$$

denote the GDP(MAP) estimator, where $\alpha'_n = 2\sigma^2(\alpha_n + 1)$ and $\eta'_n = \sigma\eta_n$. Let $\mathcal{A}_n = \{j : \beta_{nj}^{(\infty)} \neq 0, j = 1, \dots, p\}$. Suppose that $\alpha'_n \rightarrow \infty$, $\alpha'_n/\sqrt{n} \rightarrow 0$ and, $\eta'_n\sqrt{n} \rightarrow c < \infty$. Then $\beta_n^{(\infty)}$ is

1. consistent in variable selection in that $\lim_{n \rightarrow \infty} \mathbb{P}(\mathcal{A}_n = \mathcal{A}) = 1$;
2. asymptotically normal with $\sqrt{n}(\beta_{n\mathcal{A}}^{(\infty)} - \beta_{\mathcal{A}}^*) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{C}_{\mathcal{A}}^{-1})$.

Remark 3. More generally, the above results hold if $\alpha'_n/(\sqrt{n}\eta'_n) \rightarrow \infty$ and $\alpha'_n/\sqrt{n} \rightarrow 0$.

Theorem 2. Let $\beta_n^{(1)}$ denote the GDP(OS) estimator in (4.5) and $\mathcal{A}_n = \{j : \beta_{nj}^{(1)} \neq 0, j = 1, \dots, p\}$. Suppose that $\alpha_n^\dagger \rightarrow \infty$, $\alpha_n^\dagger/\sqrt{n} \rightarrow 0$, and $\eta_n^\dagger\sqrt{n} \rightarrow c < \infty$. Then $\beta_n^{(1)}$ is

1. consistent in variable selection in that $\lim_{n \rightarrow \infty} \mathbb{P}(\mathcal{A}_n = \mathcal{A}) = 1$;
2. asymptotically normal with $\sqrt{n}(\beta_{n\mathcal{A}}^{(1)} - \beta_{\mathcal{A}}^*) \xrightarrow{d} N(0, \sigma^2 \mathbf{C}_{\mathcal{A}}^{-1})$.

The proofs are deferred to Section 8.

5. Experiments

5.1. Simulation

In this section, we compare the proposed estimators to the posterior means obtained under the normal, Laplace, and horseshoe priors, to the Bayesian model averaged (BMA) estimator, as well as to the sparse estimates resulting from LASSO (Tibshirani (1996)) and SCAD (Fan and Li (2001)). GDP(PM) and GDP(MAP) denote the posterior mean and the MAP estimates, respectively, under the generalized double Pareto prior. Hyper-parameter values are provided in footnotes of Tables 5.1 and 5.2 when fixed in advance and are otherwise treated as random with the priors specified in Section 3. When not fixed, we first obtain the posterior means of the hyper-parameters from an initial Bayesian analysis, then use them in the calculation of the MAP estimates.

We generated $n = \{50, 400\}$ observations from $y_i = \mathbf{x}_i' \boldsymbol{\beta}^* + \epsilon_i$, where the x_{ij} were standard normals with $\text{Cov}(x_j, x_{j'}) = 0.5^{|j-j'|}$, $\epsilon_i \sim N(0, \sigma^2)$, and $\sigma = 3$. We used the following $\boldsymbol{\beta}^*$ configurations:

Model 1: 5 randomly chosen components of $\boldsymbol{\beta}^*$ set to 1 and the rest to 0.

Model 2: 5 randomly chosen components of $\boldsymbol{\beta}^*$ set to 3 and the rest to 0.

Model 3: 10 randomly chosen components of $\boldsymbol{\beta}^*$ set to 1 and the rest to 0.

Model 4: 10 randomly chosen components of $\boldsymbol{\beta}^*$ set to 3 and the rest to 0.

Model 5: $\boldsymbol{\beta}^* = (0.85, \dots, 0.85)'$.

In our experiments \mathbf{y} and the columns of \mathbf{X} were centered and the columns of \mathbf{X} scaled to have unit length. For the calculation of competing estimators we used `lars` (Hastie and Efron (2011)), `SIS` (Fan et al. (2010)), `monomvn` (Gramacy (2010)) and `BAS` (Clyde and Littman (2005), Clyde, Ghosh, and Littman (2010)) packages in R. We mainly followed the default settings provided by the packages. Under the normal prior, the so-called “ridge” parameter was given an inverse gamma prior with shape and scale parameters 10^{-3} . Under the Laplace prior, as a default choice, a gamma prior was placed on the “LASSO parameter” λ^2 , as given in (6) of Park and Casella (2008), with shape and rate parameters 2 and 0.1, respectively. Under the horseshoe prior, the `monomvn` package uses the hierarchy given in Section 1.1 of Carvalho, Polson, and Scott (2010). For BMA, we used the default settings of the `BAS` package that employs the Zellner-Siow prior given in Section 3.1 of Liang et al. (2008). The tuning for LASSO and SCAD were carried out by the criteria given in Yuan and Lin (2005) and Wang, Li, and Tsai (2007), respectively, avoiding cross-validation.

100 data sets were generated for each case. In Table 5.1, we report the median model error. Model error was calculated as $(\beta^* - \hat{\beta})' \mathbf{C} (\beta^* - \hat{\beta})$, where \mathbf{C} is the variance-covariance matrix that generated \mathbf{X} and $\hat{\beta}$ denotes the estimator in use. The values in the subscripts give the bootstrap standard error of the median model error values obtained. The bootstrap standard error was calculated by generating 500 bootstrap samples from 100 model error values, finding the median model error for each case, and then calculating the standard error for it. Under each model, the best three performances are boldfaced in the tables.

GDP(PM) estimates showed a similar performance to that of horseshoe under sparse setups. GDP(PM) (with α and η unknown) also showed great flexibility in adapting to dense models with small signals. GDP(MAP) estimates performed similarly to SCAD and much better than LASSO, particularly so with increasing sparsity, signal and/or sample size. The GDP(PM) and GDP(MAP) calculations are straightforward and computationally inexpensive due to the normal (and Laplace) scale mixture representation used. Being able to use a simple Gibbs sampler (especially when $\alpha = \eta = 1$) makes the procedure attractive for the average user.

Letting $\alpha = \eta = 1$ may be somewhat restrictive if the underlying model is very dense or very sparse, but in the cases we considered, it performed comparably to others and we believe that it constitutes a good default prior similar to standard Cauchy with the added advantage of thresholding ability. Although we do not take up $p \gg n$ cases in this paper, in such situations much larger values of α would need to be chosen to adjust for multiplicity.

5.2. Inferences on Hyper-parameters

Here we take a closer look at the inferences on the hyper-parameters obtained from an individual data set for Models 2 and 5 from Section 5.1. This gives us some insight into how α and η are inferred with changing sample size and sparsity structure. Note that GDP(PM)² is more restrictive than GDP(PM) as η is fixed, treating only α as unknown. Figure 5.6 gives the marginal posteriors of α and η in cases of GDP(PM)² and GDP(PM) as described in Section 5.1, while Table 5.2 reports the posterior means for α and η , as well as model error (ME) performance (as calculated in Section 5.1) on the particular data set used. We clearly observe the adaptive nature and higher flexibility of GDP(PM) moving from a sparse to a dense model with a big increase, particularly in η , flattening the prior on β . There is not quite as much wiggle room in the case of GDP(PM)². All it

Table 5.1: Model error comparisons.

$n = 50$					
Method	Model 1	Model 2	Model 3	Model 4	Model 5
Normal	2.299 _{0.085}	4.879 _{0.263}	2.585 _{0.134}	4.972 _{0.385}	2.886 _{0.150}
Laplace	2.634 _{0.137}	3.662 _{0.233}	2.837 _{0.126}	4.326 _{0.211}	3.458 _{0.120}
Horseshoe	2.264 _{0.086}	2.316 _{0.167}	3.205 _{0.140}	3.929 _{0.218}	4.409 _{0.130}
BMA	2.451 _{0.123}	1.647 _{0.126}	4.043 _{0.233}	3.062 _{0.194}	6.015 _{0.301}
GDP(PM) ¹	2.306 _{0.114}	2.405 _{0.192}	3.193 _{0.215}	4.123 _{0.304}	4.283 _{0.142}
GDP(PM) ²	2.303 _{0.095}	2.309 _{0.195}	3.124 _{0.153}	3.910 _{0.237}	4.451 _{0.109}
GDP(PM)	2.271 _{0.085}	2.606 _{0.167}	3.047 _{0.147}	4.348 _{0.171}	3.640 _{0.134}
GDP(MAP) ¹	3.414 _{0.148}	1.619 _{0.150}	5.605 _{0.298}	2.970 _{0.168}	8.769 _{0.403}
GDP(MAP) ²	4.250 _{0.354}	1.618 _{0.153}	6.331 _{0.300}	3.040 _{0.163}	9.308 _{0.377}
GDP(MAP)	4.876 _{0.355}	2.091 _{0.182}	4.299 _{0.222}	3.740 _{0.284}	5.724 _{0.177}
LASSO	2.183 _{0.124}	2.618 _{0.152}	3.258 _{0.194}	3.531 _{0.172}	5.646 _{0.229}
SCAD	3.732 _{0.214}	2.132 _{0.229}	5.249 _{0.239}	3.179 _{0.193}	8.505 _{0.387}
$n = 400$					
Normal	0.395 _{0.014}	0.455 _{0.019}	0.426 _{0.016}	0.455 _{0.024}	0.412 _{0.013}
Laplace	0.315 _{0.016}	0.374 _{0.014}	0.388 _{0.016}	0.422 _{0.015}	0.457 _{0.014}
Horseshoe	0.219 _{0.016}	0.205 _{0.010}	0.341 _{0.014}	0.346 _{0.009}	0.514 _{0.023}
BMA	0.151 _{0.011}	0.125 _{0.005}	0.240 _{0.016}	0.211 _{0.009}	0.646 _{0.037}
GDP(PM) ¹	0.233 _{0.016}	0.206 _{0.009}	0.326 _{0.015}	0.284 _{0.014}	0.625 _{0.031}
GDP(PM) ²	0.228 _{0.017}	0.215 _{0.009}	0.332 _{0.013}	0.303 _{0.010}	0.579 _{0.027}
GDP(PM)	0.248 _{0.017}	0.182 _{0.007}	0.377 _{0.016}	0.362 _{0.012}	0.466 _{0.016}
GDP(MAP) ¹	0.154 _{0.014}	0.111 _{0.011}	0.286 _{0.016}	0.210 _{0.011}	0.739 _{0.043}
GDP(MAP) ²	0.161 _{0.013}	0.111 _{0.010}	0.284 _{0.016}	0.210 _{0.009}	0.652 _{0.035}
GDP(MAP)	0.185 _{0.017}	0.119 _{0.010}	0.326 _{0.016}	0.336 _{0.010}	0.478 _{0.020}
LASSO	0.251 _{0.014}	0.276 _{0.014}	0.339 _{0.020}	0.348 _{0.011}	0.485 _{0.021}
SCAD	0.121 _{0.010}	0.118 _{0.008}	0.233 _{0.011}	0.206 _{0.017}	0.469 _{0.019}

¹ $\alpha = 1, \eta = 1$; ² $\eta = 1$

can do is to drive α smaller to allow heavier tails to accommodate a dense structure. As observed in Table 5.1, however, GDP(PM)² performs comparably in sparse cases.

Table 5.2: Posterior means of the hyper-parameters and the resulting model error.

$n = 50$			$n = 400$		
		GDP(PM)	GDP(PM) ²	GDP(PM)	GDP(PM) ²
Model 2	α	2.464	1.165	0.688	0.870
	η	4.181	—	0.614	—
	ME	2.443	2.219	0.149	0.181
Model 5	α	5.262	1.200	9.400	0.560
	η	9.476	—	51.735	—
	ME	6.290	7.019	0.518	0.614

² $\eta = 1$

6. Data Example

We consider the ozone data analyzed by Breiman and Friedman (1985) and by Casella and Moreno (2006). The original data set contains 13 variables and 366 observations. The modeled response is the daily maximum one-hour averaged ozone reading in Los Angeles over 330 days in 1976. There are $p = 12$ predictors considered and deleting incomplete observations leaves $n = 203$ observations. For validation, the data were split into a training set containing 180 observations and a test set containing 23 observations. We considered models including main effects, quadratic, and two-way interaction terms resulting in 2^{90} possible subsets. The complex correlation structure of the data is illustrated in Figure 6.7.

Figure 6.8 summarizes the performance of the proposed estimators and their competitors. Median values for R_{test}^2 and the ± 2 standard error intervals were obtained by running the methods on 100 different random training-test splits. Standard errors were computed via bootstrapping the medians 500 times.

The median number of predictors retained in the model by all three GDP(MAP) estimates was only 4 while it was 14 and 9 for LASSO and SCAD. Hence GDP(MAP) promoted much sparser models. In terms of prediction, GDP(PM)¹ yielded the second best results after BMA, with GDP(PM)², GDP(PM), and the horseshoe estimator all having somewhat worse performance. These shrinkage priors are designed to mimic model averaging behavior, so we expected to obtain results that were competitive with, but not better than, BMA. The improved performance for GDP(PM)¹ may be attributed to the use of default hyper-parameter values that were fixed in advance at values thought to produce good performance in sparse settings. Treating the hyper-parameters as unknown is appealing from the standpoint of flexibility, but in practice the data may not inform sufficiently about their values to outperform a good default choice. GDP(MAP)¹ and SCAD both performed within the standard error range of LASSO, while retaining a smaller number of variables in the model. As it is important to account for model uncertainty in prediction, the posterior mean estimator under the GDP prior is appealing in mimicking BMA. In addition, obtaining a simple model containing a relatively small number of predictors is often important, since such models are more likely to be used in fields in which predictive black boxes are not acceptable and practitioners desire interpretable predictive models.

7. Discussion

We have proposed a hierarchical prior obtained through a particular scale mixture of normals where the resulting marginal prior has a folded generalized Pareto density thresholded at zero. This prior combines the best of both worlds in that fully Bayes inferences are feasible through its hierarchical representation, providing a measure of uncertainty in estimation, while the resulting marginal prior on the regression coefficients induces a penalty function that allows for the analysis of frequentist properties under maximum a posteriori estimation. The resulting posterior mean estimator can be argued to be mimicking a Bayesian model averaging behavior through mixing over higher level hyper-parameters. Although Bayesian model averaging is appealing, it can be argued that allowing parameters to be arbitrarily close to zero instead of exactly equal to zero may be more natural in some problems. Hence we have a procedure that not only bridges two paradigms – Bayesian shrinkage estimation and regularization – but also yields three useful tools: a sparse estimator with good frequentist properties through maximum a posteriori estimation, a posterior mean estimator that mimics a model averaging behavior, and a useful measure of uncertainty around the observed estimates. In addition, the proposed methods have substantial computational advantages in relying on simple block-updated Gibbs sampling, while BMA requires sampling from a model space with 2^p models. Given the simple and fast computation and the excellent performance in small sample simulation studies, the generalized double Pareto should be useful as a shrinkage prior in a broad variety of Bayesian hierarchical models, while also suggesting close relationships with frequentist penalized likelihood approaches. The proposed prior can be used in generalized linear models, shrinkage of basis coefficients in nonparametric regression, and in such settings as factor analysis and nonparametric Bayes modeling.

8. Technical Details

Proof of Theorem 1. The proof follows along similar lines as does the proof of Theorem 2 in Zou (2006). We first prove asymptotic normality. Let $\beta = \beta^* + \mathbf{u}/\sqrt{n}$ and

$$V_n(\mathbf{u}) = \left\{ \mathbf{y} - \sum_{j=1}^p \mathbf{x}_j \left(\beta_j^* + \frac{u_j}{\sqrt{n}} \right) \right\}^2 + \alpha'_n \sum_{j=1}^p \log \left\{ \left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| + \eta'_n \right\}.$$

Let $\hat{\mathbf{u}}_n = \arg \min V_n(\mathbf{u})$, suggesting $\hat{\mathbf{u}}_n = \sqrt{n}(\beta_n^{(\infty)} - \beta^*)$. Now

$$V_n(\mathbf{u}) - V_n(\mathbf{0}) = \mathbf{u}' \left(\frac{1}{n} \mathbf{X}' \mathbf{X} \right) \mathbf{u} - 2 \frac{\boldsymbol{\epsilon}' \mathbf{X}}{\sqrt{n}} \mathbf{u} + \alpha'_n \sum_{j=1}^p \log \frac{|\beta_j^* + u_j/\sqrt{n}| + \eta'_n}{|\beta_j^*| + \eta'_n},$$

and we know that $\mathbf{X}' \mathbf{X}/n \rightarrow \mathbf{C}$ and $\boldsymbol{\epsilon}' \mathbf{X}/\sqrt{n} \xrightarrow{d} \mathbf{W} \stackrel{d}{=} \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{C})$. Consider the limiting behavior of the third term, noting that $\lim_{a \rightarrow \infty} (1 + b/a)^a = e^b$. If $\beta_j^* \neq 0$, then $\alpha'_n \log \frac{|\beta_j^* + u_j/\sqrt{n}| + \eta'_n}{|\beta_j^*| + \eta'_n} \leq \alpha'_n \log \frac{|\beta_j^*| + |u_j/\sqrt{n}| + \eta'_n}{|\beta_j^*| + \eta'_n} = \alpha'_n \log \left(1 + \frac{|u_j/\sqrt{n}|}{|\beta_j^*| + \eta'_n} \right) \rightarrow 0$. If $\beta_j^* = 0$, then $\alpha'_n \log \frac{|u_j/\sqrt{n}| + \eta'_n}{\eta'_n} = \alpha'_n \log \left(1 + \frac{|u_j/\sqrt{n}|}{\eta'_n} \right)$ which is 0 if $u_j = 0$, and diverges otherwise. By Slutsky's Theorem

$$V_n(\mathbf{u}) - V_n(\mathbf{0}) \xrightarrow{d} \begin{cases} \mathbf{u}'_{\mathcal{A}} \mathbf{C}_{\mathcal{A}} \mathbf{u}_{\mathcal{A}} - 2 \mathbf{u}'_{\mathcal{A}} \mathbf{W}_{\mathcal{A}} & \text{if } u_j = 0 \ \forall j \notin \mathcal{A} \\ \infty & \text{otherwise.} \end{cases}$$

$V_n(\mathbf{u}) - V_n(\mathbf{0})$ is convex and the unique minimum of the right hand side is $(\mathbf{C}_{\mathcal{A}}^{-1} \mathbf{W}_{\mathcal{A}}, \mathbf{0})'$. By epiconvergence (Geyer (1994), Knight and Fu (2000)),

$$\hat{\mathbf{u}}_{n\mathcal{A}} \xrightarrow{d} \mathbf{C}_{\mathcal{A}}^{-1} \mathbf{W}_{\mathcal{A}}, \quad \hat{\mathbf{u}}_{n\mathcal{A}^c} \xrightarrow{d} \mathbf{0}. \quad (8.1)$$

Since $\mathbf{W}_{\mathcal{A}} \stackrel{d}{=} \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{C}_{\mathcal{A}})$, this proves asymptotic normality.

Now $\forall j \in \mathcal{A}$, $\beta_{nj}^{(\infty)} \xrightarrow{p} \beta_j^*$; thus $\mathbb{P}(j \in \mathcal{A}_n) \rightarrow 1$. Hence for consistency, it is sufficient to show that $\forall j' \notin \mathcal{A}$, $\mathbb{P}(j' \in \mathcal{A}_n) \rightarrow 0$. Consider the event $j' \in \mathcal{A}_n$. By the KKT optimality conditions, $2\mathbf{x}'_{j'}(\mathbf{y} - \mathbf{X}\beta_n^{(\infty)}) = \frac{\alpha'_n}{\eta'_n + |\beta_{nj'}^{(\infty)}|}$. Noting that $\sqrt{n}\beta_{nj'}^{(\infty)} \xrightarrow{p} 0$ by (8.1), $\frac{\alpha'_n}{\sqrt{n}\eta'_n + \sqrt{n}|\beta_{nj'}^{(\infty)}|} \rightarrow \infty$, while

$$\frac{2\mathbf{x}'_{j'}(\mathbf{y} - \mathbf{X}\beta_n^{(\infty)})}{\sqrt{n}} = 2 \left\{ \frac{\mathbf{x}'_{j'} \mathbf{X} \sqrt{n}(\beta^* - \beta_n^{(\infty)})}{n} + \frac{\mathbf{x}'_{j'} \boldsymbol{\epsilon}}{\sqrt{n}} \right\}.$$

By (8.1) and Slutsky's Theorem, we know that both terms in the brackets converge in distribution to some normal, so

$$\mathbb{P}(j' \in \mathcal{A}_n) \leq \mathbb{P} \left\{ 2\mathbf{x}'_{j'}(\mathbf{y} - \mathbf{X}\beta_n^{(\infty)}) = \frac{\alpha'_n}{|\beta_{nj'}^{(\infty)}| + \eta'_n} \right\} \rightarrow 0.$$

This concludes the proof. \square

Proof of Theorem 2. We modify the proof of Theorem 2 in Zou (2006). Here $\beta_n^{(0)}$ denotes the least squares estimator. We first prove asymptotic normality. Let $\beta = \beta_n^* + \mathbf{u}/\sqrt{n}$ and

$$V_n(\mathbf{u}) = \left\{ y - \sum_{j=1}^p x_j \left(\beta_{nj}^* + \frac{u_j}{\sqrt{n}} \right) \right\}^2 + \alpha_n^\dagger \sum_{j=1}^p \left| \beta_{nj}^* + \frac{u_j}{\sqrt{n}} \right| \left(|\beta_{nj}^{(0)}| + \eta_n^\dagger \right)^{-1}.$$

Let $\hat{\mathbf{u}}_n = \arg \min V_n(\mathbf{u})$, suggesting $\hat{\mathbf{u}}_n = \sqrt{n}(\beta_n^{(1)} - \beta_n^*)$. Now

$$\begin{aligned} V_n(\mathbf{u}) - V_n(\mathbf{0}) &= \mathbf{u}' \left(\frac{1}{n} \mathbf{X}' \mathbf{X} \right) \mathbf{u} - 2 \frac{\boldsymbol{\epsilon}' \mathbf{X}}{\sqrt{n}} \mathbf{u} \\ &\quad + \frac{\alpha_n^\dagger}{\sqrt{n}} \sum_{j=1}^p \left(|\beta_{nj}^{(0)}| + \eta_n^\dagger \right)^{-1} \sqrt{n} \left(\left| \beta_{nj}^* + \frac{u_j}{\sqrt{n}} \right| - |\beta_{nj}^*| \right), \end{aligned}$$

and we know that $\mathbf{X}' \mathbf{X}/n \rightarrow \mathbf{C}$ and $\boldsymbol{\epsilon}' \mathbf{X}/\sqrt{n} \xrightarrow{d} \mathbf{W} \stackrel{d}{=} \mathbf{N}(0, \sigma^2 \mathbf{C})$. Consider the limiting behavior of the third term. If $\beta_{nj}^* \neq 0$ then, by the Continuous Mapping Theorem, $\{|\beta_{nj}^{(0)}| + \eta_n^\dagger\}^{-1} \xrightarrow{p} \{|\beta_{nj}^*| + \eta_n^\dagger\}^{-1}$ and $\sqrt{n}(|\beta_{nj}^* + u_j/\sqrt{n}| - |\beta_{nj}^*|) \rightarrow u_j \text{sgn}(\beta_{nj}^*)$. By Slutsky's Theorem, $(\alpha_n^\dagger/\sqrt{n})\{|\beta_{nj}^{(0)}| + \eta_n^\dagger\}^{-1} \sqrt{n}(|\beta_{nj}^* + u_j/\sqrt{n}| - |\beta_{nj}^*|) \xrightarrow{p} 0$. If $\beta_{nj}^* = 0$, then $\sqrt{n}(|\beta_{nj}^* + u_j/\sqrt{n}| - |\beta_{nj}^*|) = |u_j|$ and $\alpha_n^\dagger\{|\beta_{nj}^{(0)}| + \eta_n^\dagger\}^{-1}/\sqrt{n} = \alpha_n^\dagger/(\sqrt{n}|\beta_{nj}^{(0)}| + \sqrt{n}\eta_n^\dagger)$, where $\sqrt{n}\beta_{nj}^{(0)} = O_p(1)$. Again by Slutsky's Theorem,

$$V_n(\mathbf{u}) - V_n(\mathbf{0}) \xrightarrow{d} \begin{cases} \mathbf{u}'_{\mathcal{A}} \mathbf{C}_{\mathcal{A}} \mathbf{u}_{\mathcal{A}} - 2 \mathbf{u}'_{\mathcal{A}} \mathbf{W}_{\mathcal{A}} & \text{if } u_j = 0 \text{ for all } j \notin \mathcal{A} \\ \infty & \text{otherwise.} \end{cases}$$

$V_n(\mathbf{u}) - V_n(\mathbf{0})$ is convex and the unique minimum of the right hand side is $(\mathbf{C}_{\mathcal{A}}^{-1} \mathbf{W}_{\mathcal{A}}, \mathbf{0})'$. By epiconvergence (Geyer (1994), Knight and Fu (2000)),

$$\hat{\mathbf{u}}_{n\mathcal{A}} \xrightarrow{d} \mathbf{C}_{\mathcal{A}}^{-1} \mathbf{W}_{\mathcal{A}}, \quad \hat{\mathbf{u}}_{n\mathcal{A}^c} \xrightarrow{d} \mathbf{0}. \quad (8.2)$$

Since $\mathbf{W}_{\mathcal{A}} \stackrel{d}{=} \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{C}_{\mathcal{A}})$, this proves the asymptotic normality.

Now $\forall j \in \mathcal{A}$, $\beta_{nj}^{(1)} \xrightarrow{p} \beta_{nj}^*$; thus $\mathbb{P}(j \in \mathcal{A}_n) \rightarrow 1$. We show that for all $j' \notin \mathcal{A}$, $\mathbb{P}(j' \in \mathcal{A}_n) \rightarrow 0$. Consider the event $j' \in \mathcal{A}_n$. By the KKT optimality conditions, $2\mathbf{x}'_{j'}(\mathbf{y} - \mathbf{X}\beta_n^{(1)}) = \alpha_n^\dagger(|\beta_{nj'}^{(0)}| + \eta_n^\dagger)^{-1}$. We know that $\alpha_n^\dagger(|\beta_{nj'}^{(0)}| + \eta_n^\dagger)^{-1}/\sqrt{n} \xrightarrow{p} \infty$,

while

$$\frac{2\mathbf{x}'_{j'}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n^{(1)})}{\sqrt{n}} = 2 \left\{ \frac{\mathbf{x}'_{j'}\mathbf{X}\sqrt{n}(\boldsymbol{\beta}_n^* - \boldsymbol{\beta}_n^{(1)})}{n} + \frac{\mathbf{x}'_{j'}\boldsymbol{\epsilon}}{\sqrt{n}} \right\}.$$

By (8.2) and Slutsky's Theorem, we know that both terms in the brackets converge in distribution to some normal, so

$$\mathbb{P}(j' \in \mathcal{A}_n) \leq \mathbb{P} \left\{ 2\mathbf{x}'_{j'}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n^{(1)}) = \frac{\alpha_n^\dagger}{|\beta_{nj'}^{(0)}| + \eta_n^\dagger} \right\} \rightarrow 0,$$

which proves consistency. \square

Acknowledgments

This work was supported by Award Number R01ES017436 from the National Institute of Environmental Health Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Environmental Health Sciences or the National Institutes of Health. Jaeyong Lee was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (20110027353).

References

- Berger, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics* **8**, 716–761.
- Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York.
- Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* **24**, 2350–2383.
- Breiman, L. and Friedman, J. H. (1985), Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* **80**.
- Candes, E. J., Wakin, M. B., and Boyd, S. P. (2008). Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications* **14**, 877–905.

- Carvalho, C., Polson, N., and Scott, J. (2009). Handling sparsity via the horseshoe. *JMLR: W&CP* **5**.
- Carvalho, C., Polson, N., and Scott, J. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480.
- Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection. *Journal of the American Statistical Association* **101**.
- Cevher, V. (2009). Learning with compressible priors. *Advances in Neural Information Processing Systems* **22**.
- Clyde, M., Ghosh, J., and Littman, M. L. (2010). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics* **20**, 80–101.
- Clyde, M. and Littman, M. (2005). *Bayesian model averaging using Bayesian adaptive sampling – BAS package manual*.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32**, 407–499.
- Fan, J., Feng, Y., Samworth, R., and Wu, Y. (2010). *SIS package manual*.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.
- Figueiredo, M. A. T. (2003). Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**, 1150–1159.
- Fu, W. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**, 397–416.
- Garrigues, P. J. (2009). Sparse coding models of natural images: Algorithms for efficient inference and learning of higher-order structure. PhD Thesis, University of California, Berkeley.
- Geyer, C. J. (1994). On the asymptotics of constrained M-estimation. *The Annals of Statistics* **22**, 1993–2010.

- Gramacy, R. B. (2010). *Estimation for multivariate normal and Student-t data with monotone missingness – Monomvn package manual*.
- Griffin, J. E. and Brown, P. J. (2007). Bayesian adaptive lassos with non-convex penalization. Technical Report.
- Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* **5**, 171–188.
- Hans, C. (2009). Bayesian lasso regression. *Biometrika* **96**, 835–845.
- Hastie, T. and Efron, B. (2011). *Lars package manual*.
- Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics* **28**, 1356–1378.
- Lee, A., Caron, F., Doucet, A., and Holmes, C. (2012). Bayesian sparsity-path-analysis of genetic association signal using generalized t priors. *Statistical Applications in Genetics and Molecular Biology* **11**.
- Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association* **103**, 410–423.
- McDonald, J. B. and Newey, W. K. (1988). Partially adaptive estimation of regression models via the generalized t distribution. *Econometric Theory* **4**, 428–457.
- Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association* **103**, 681–686.
- Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics* **3**, 119–131.
- Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics 9*. Oxford University Press.
- Ritter, C. and Tanner, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the griddy-Gibbs sampler. *Journal of the American Statistical Association* **97**, 861–868.

- Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics* **42**, 385–388.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* **58**, 267–288.
- Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* **1**.
- Wang, H., Li, R., and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.
- West, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74**, 646–648.
- Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association* **100**, 1215–1225.
- Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research* **7**.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.
- Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* **36**, 1509–1533.

SAS Institute Inc., Durham, NC 27513, USA

E-mail: artin.armagan@sas.com

Department of Statistical Science, Duke University, Durham, NC 27708, USA

E-mail: dunson@stat.duke.edu

Department of Statistics, Seoul National University, Seoul, 151-747, Korea

E-mail: leejyc@gmail.com

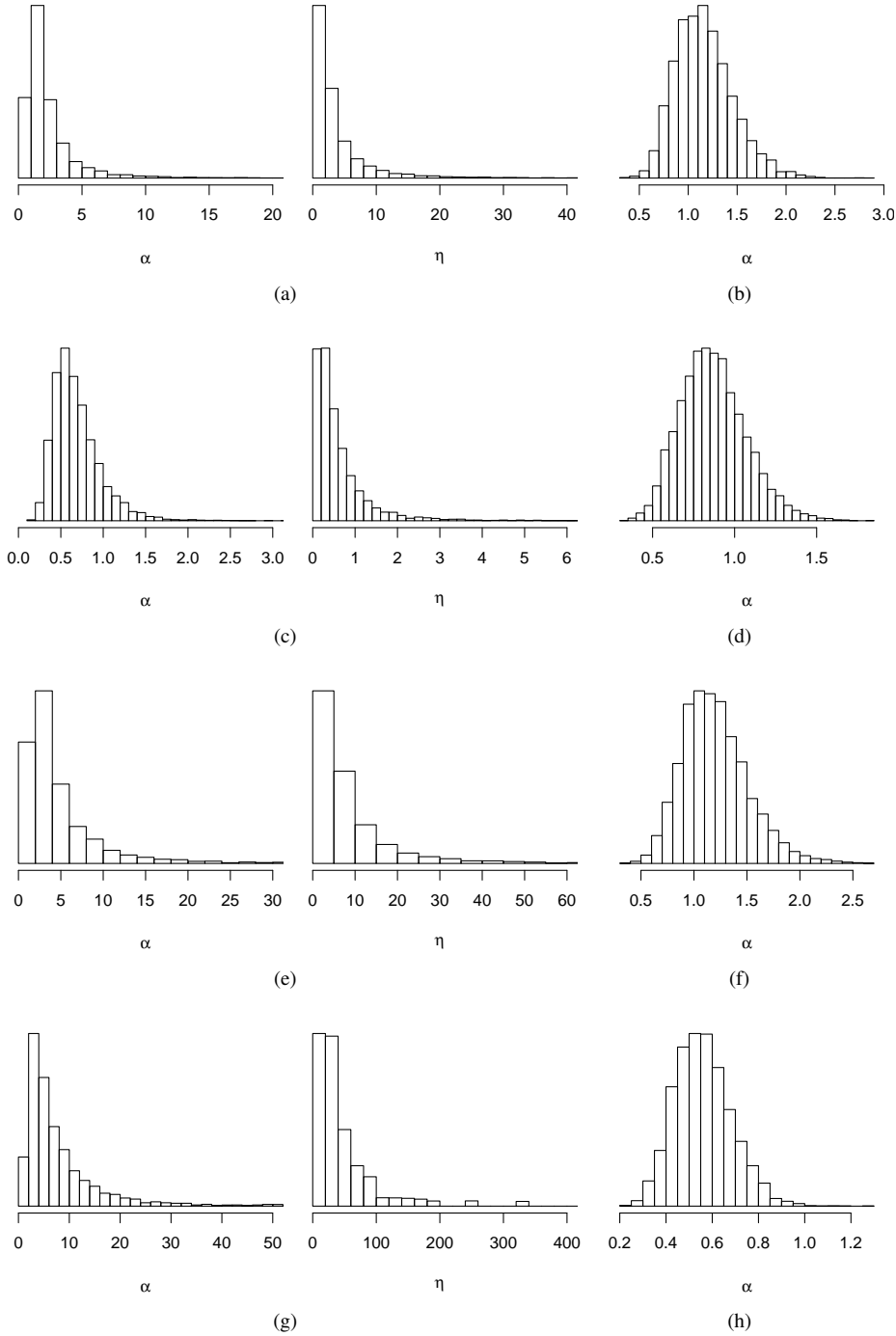


Figure 5.6: Inferences for (a) GDP(PM) for $n = 50$ under Model 2, (b) GDP(PM)^2 for $n = 50$ under Model 2, (c) GDP(PM) for $n = 400$ under Model 2, (d) GDP(PM)^2 for $n = 400$ under Model 2, (e) GDP(PM) for $n = 50$ under Model 5, (f) GDP(PM)^2 for $n = 50$ under Model 5, (g) GDP(PM) for $n = 400$ under Model 5, (h) GDP(PM)^2 for $n = 400$ under Model 2 .

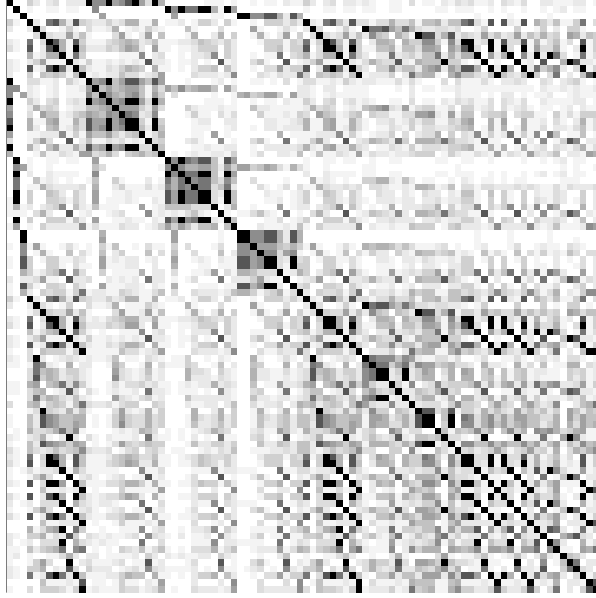
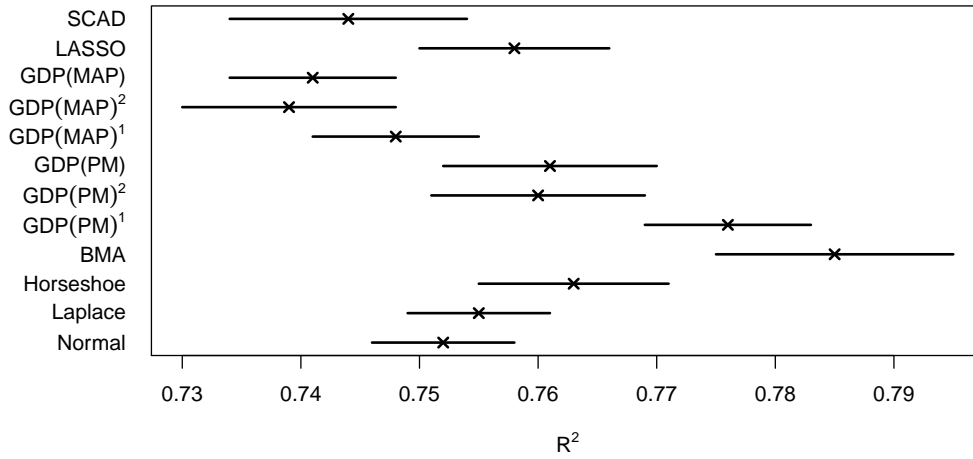


Figure 6.7: The correlation structure of the Ozone data.

Figure 6.8: Out-of-sample performance comparisons for Ozone data. (\times) denotes the median value for R^2_{test} while the lines represent the ± 2 standard error regions. ¹ $\alpha = 1, \eta = 1$; ² $\eta = 1$.